

Inhalt

- Einfuehrung**
 - Geschichte, Zielgruppe
 - Z39.50
 - Struktuierte Daten
 - verteilte Suche

- Detail**
 - Normierung
 - Dublettenerkennung
 - parallele Suche
 - Fazit, Grenzen des Systems

- Vorfuehrung**
 - Import von strukturierten Daten
 - Verteilte Suche

Zack - eine Metasuchmaschine fuer Bibliotheken

Berliner Bibliothekswissenschaftliches
Kolloquium, Humboldt Universitaet Berlin,
9-Mai-2000.

by Wolfram Schneider
wolfram@schneider.org
<http://wolfram.schneider.org>

Einleitung

- Geschichte**
 - Projekt KOBV am ZIB, von 1997 bis 2000
 - Kooperativer Bibliotheksverbund
 - Bedarf an Dienstleistungen sofort

- Aufgabe**
 - Web-Interface zur Datenbank ILTIS der Deutschen Bibliothek fuer Copy-Cataloging in Cottbus und Frankfurt/O.
 - Fremddaten (DNB)

- Zielgruppe**
 - Katalogisierer, Bibliothekare, Retrokatalogisierung
 - Virtueller Verbund

Zusammenfassung

- Zack ist eine Suchmaschine fuer Bibliotheksdatenbanken die ueber das Z39.50 Protokoll ansprechbar sind.

- Das Ergebnis sind strukturierte Daten (MAB2), die in das eigene Bibliotheksystem uebernommen werden koennen.

- Bei der verteilten Suche wird gleichzeitig in mehreren Datenbanken gesucht. Dubletten werden als solche erkannt.

WWW-Z39.50 Gateway

Vorteile von Z39.50 mit Vorteilen des Web verbinden

Vorteile fuer die Benutzer

- Web-Browser weit verbreitet
- gewohnte Software, einfach zu bedienen
- keine zusaetzlichen Kosten fuer Z39.50 Clients
- keine Installation und Wartung von Z39.50 Clients

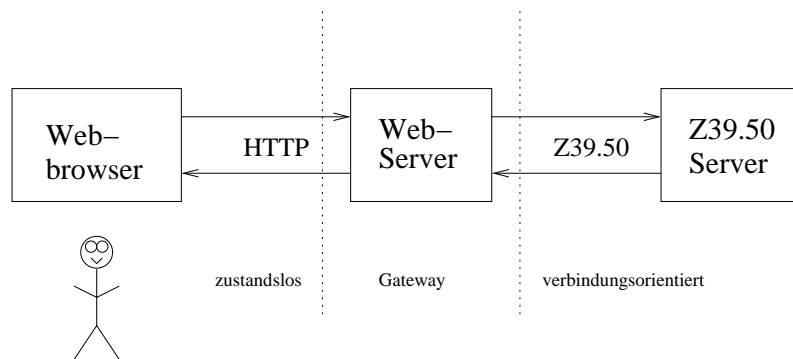
Nachteile

- eingeschaernte Funktionalitaet
- Performance

Z39.50 Allgemein

- Kommunikationsprotokoll, ISO23950: Information Retrieval
- Client/Server System
- Datenbank-Esperanto: mit jedem Client einen Dialog mit jeder Datenbank fuehren koennen
- unabhangig von Datenbank, lokale Anfragesyntax, Hersteller
- seit 1984 entwickelt, Library of Congress

WWW-Z39.50 Gateway Modell



Z39.50 Detail

- verbindungsorientiert (session/Sitzung)
- wichtigste Services
 - init Verbindungsaufbau
 - search Suche
 - present Ausgabe der Datensätze
 - close Verbindungsabbau
- weitere Services
 - scan Registersuche
 - sort Sortieroptionen fuer Ergebnisse
 - explain (Target profile)

Normierung

was ist das: unterschiedliche Schreibweisen vereinheitlichen die keine inhaltliche Bedeutung haben

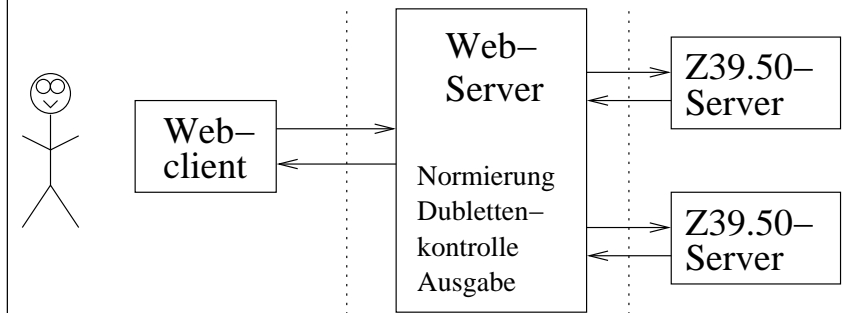
warum notwendig

- Daten unterschiedlicher Herkunft
- Zeichensätze
- Tippfehler
- unterschiedliche Katalogisierungsregeln

Beispiele

- Autor: "Dalitz, Wolfgang" -> "dalitz, wolfgang"
- ISBN: "ISBN 3-928861-23-9" -> "3928861239"

WWW-Z39.50 Gateway Verteilte Suche



Dublettenerkennung

- Ziel: gleiche oder ähnliche Datensätze erkennen
- nicht besser als ein Mensch
- Expertensystem (MYCIN)

- gleich: gleiche Attribute
- ähnlich: kleinere Abweichungen ignorieren
 - Tippfehler
 - +/- Seitenzahlen

Parallele Suche

- gleichzeitig Verbindungen aufbauen
- gleichzeitig suchen
- gleichzeitig Lesen
- max. 5-10 Datensätze pro Sekunde laden
- z.Z. bis zu 14 Datenbanken, insgesamt in 40 Mio Datensätzen

Bibliotheksverbund Bayern, KOBV Berlin-Brandenburg, Gemeinsamer Bibliotheksverbund, TU Braunschweig, FH Potsdam, Uni Potsdam, Uni Duesseldorf, FH Brandenburg, Oeffentl. Bibl. Berlins, EUV Frankfurt/O, BTU Cottbus, Max-Planck-Inst. f. Bildungsforschung, Die Deutsche Bibliothek, Suedwestdeutscher Bibliotheksverbund

Probleme - Technik

Probleme laufenden Betrieb - Technik

- instabile Z39.50 Server
- Betasoftware Z39.50 Server
- falsch konfigurierte Z39.50 Server
- keine Dokumentation
- nur wenige Server zum Testen
- hoher Wartungsaufwand

Dublettenerkennung Expertensystem

- gewichteter Vergleich der Attribute
 - hohes Gewicht fuer Titel, Autor, ISBN
 - geringes Gewicht fuer Verlag, Jahr etc.
- Argumente, die fuer eine Dublette sprechen
- Argumente, die gegen eine Dublette sprechen
- positiver/negativer Schwellwert

Probleme - Semantik Anfragen

Semantik von Z39.50 Anfragen

- Autorsuche
 - "wolfgang dalitz"
 - "dalitz, wolfgang"
 - "dalitz,wolfgang"
- ISBN Nummer
 - ISBN Nummer mit oder ohne Bindestriche (Normierung)

Dublettenerkennung - Aufwand

- Aufwand: $n*(n-1)/2$
 - 10 Datensätze -> 45 Vergleiche
 - 30 Datensätze -> 435 Vergleiche
- Performance
 - Optimierung mit temporaerem Index
 - Cluster von aehnlichen Datensätzen
- Grenzen
 - wieviel Toleranz erlaubt

Nutzer des Systems

fuer die taegliche Arbeit beim Katalogisieren

- BTU Cottbus
- EUV Frankfurt/Oder
- Uni Potsdam
- Uni Bonn

(werk-)taeglich ca. 500-800 Suchanfragen

Einsatz der Software bzw. Teilen der Software

- ZIB/KOBV Zentrale
- TU Braunschweig
- Hochschulbibliothekszentrum NRW (HBZ)
- Universitaet des Saarlandes (SWB)

Probleme - Semantik Indexierung

Semantik Indexierung der lokalen Datenbanken

- Wortindex
- Wortgruppen (Phrasen)

- Welche Attribute fuer Indexierung
 - Schlagwoerter im Titindex
 - Autor: Buecher von oder auch ueber einen Autor
- begrenzte Zahl von Indexen
- Norm- und Titeldaten in einem Index

Wunschliste Ausgabe

- Benutzerfreundlicheres Design
- Text-Ausgabe der Treffer besser formatieren
- Super-Merged-Record: Datensaeetze mischen (bessere Schlagwoerter)
- XML Export der Datensaeetze
- Zeichensatz: nicht nur latin1

Probleme - Austauschformate

- Z39.50 orientiert sich an USMARC
- MAB2 kaum unterstuetzt und dokumentiert, z.B Kurzform
- MAB2 Hierarchien sind in Z39.50 nicht vorgesehen
- MAB2 wird unterschiedlich interpretiert
 - Beispiel: zwei ISBN Nummern - zwei Felder?
- Regelwerke
 - RAK, RAK-WB

Fazit - Grenzen des Systems

inhaltlich:

- nur so gut wie die Bibliotheksdatenbanken
- nur so gut wie die Z39.50 Server
- MAB2 Hierarchien nur eingeschaenkt

technisch:

- max. 50 Datenbanken gleichzeitig
- max. 500 Datensaeetze bei Dublettenkontrolle
- nur so schnell wie der langsamste Z39.50 Server
- geschaeetzt 10.000 Anfragen/Tag

Wunschliste Performance

- parallele Dublettenkontrolle auf Multiprozessorsystemen
- Dublettenkontrolle im Hintergrund mit unvollstaendigen Daten
- Dublettenkontrolle mit USMARC oder UNIMARC
- Aufnahme weiterer Z39.50 Server
- stehende Verbindungen zu Z39.50 Servern

Software

CGI-Scripte, Module:

- Perl5

Z395.50:

- YAZ Toolkit

Sonstiges

- CVS, recode, Apache, LaTeX, Magicpoint, xfig, xv

Betriebssystem & Hardware:

- Solaris, FreeBSD
- mind. 32MB RAM, 100MHz CPU

Fazit - was geht

Fazit verteilte Suche mit Dublettenkontrolle

- Verteilte Suche ist moeglich
- Antwortzeiten sind akzeptabel (<10 Sekunden)
- bessere Ergebnisse als bei der Suche in nur einer Datenbank
- Kurztrefferliste ist kuerzer (30-50%) und uebersichtlicher

Diplomarbeit

Wolfram Schneider: Ein verteiltes Bibliotheks-
Informationssystem auf Basis des Z39.50 Protokolls.
Juli 1999.

Eingereicht als Diplomarbeit an der TU Berlin,
Fachbereich Informatik.

Betreuer:

Josef Willenborg (ZIB)

Gutachter:

Prof. Dr. Erhard Konrad (TU Berlin)

Prof. Dr. Martin Groetschel (TU Berlin/ZIB)

Literatur

Online Demoversion von Zack, TU Braunschweig
<http://www.biblio.tu-bs.de/zack/>

Diplomarbeit in HTML, PostScript, PDF, ASCII, PalmDoc
<http://wolfram.schneider.org/lv/diplom/>

Diplomarbeit auf Papier: ZIB Preprint SC 99-21. ISSN
0933-7911.